

---

# CLASSICAL AND QUANTUM NOTIONS OF ENTROPY

---

from *Quantum Shannon Theory*, Wilde

Francis Kyungmin Park

# Contents

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Preface</b>   | <b>3</b>  |
| <b>2</b> | <b>Entropy of a Random Variable</b>                                  | <b>3</b>  |
| 2.1      | Binary Entropy Function . . . . .                                    | 4         |
| 2.2      | Properties of Entropy . . . . .                                      | 4         |
| <b>3</b> | <b>Conditional Entropy</b>   | <b>5</b>  |
| <b>4</b> | <b>Joint Entropy</b>   | <b>6</b>  |
| 4.1      | Joint Entropy Exercises . . . . .                                    | 6         |
| 4.1.1    | Exercise 10.3.1 . . . . .  | 6         |
| 4.1.2    | Exercise 10.3.2 . . . . .  | 7         |
| 4.1.3    | Exercise 10.3.3 . . . . .  | 7         |
| 4.1.4    | Exercise 10.3.4 . . . . .  | 8         |
| <b>5</b> | <b>Mutual Information</b>  | <b>8</b>  |
| 5.1      | Mutual Information Exercises . . . . .                               | 9         |
| 5.1.1    | Exercise 10.4.1 (problem statement omitted due to spacing reasons) . | 9         |
| <b>6</b> | <b>Relative Entropy</b>  | <b>10</b> |
| <b>7</b> | <b>Conditional Mutual Information</b>                                | <b>11</b> |
| <b>8</b> | <b>Entropy Inequalities</b>  | <b>11</b> |
| 8.1      | Non-Negativity of Relative Entropy . . . . .                         | 11        |
| 8.2      | Data-Processing Inequality . . . . .                                 | 12        |
| <b>9</b> | <b>Near Saturation of Entropy Inequalities</b>                       | <b>12</b> |

# 1 Preface

I do not claim any ownership of the content below; I organized it mainly for self study purposes and added some of my own commentary in between. These notes follow directly QIT by Wilde. I've included my solutions to the exercises in the book as well.

## 2 Entropy of a Random Variable

Let us consider a random variable  $X$ , which we can use to model an experiment, and the realizations  $x$  of  $X$  belong to an alphabet, which we will denote by  $\chi$ .  $p_X(x)$  refers to the probability density function of  $X$ . We denote the information content  $i(x)$  of a certain realization as follows:

$$i(x) \equiv -\log(p_X(x)) \quad (1)$$

It is standard in both classical and quantum information theory to assume the logarithm is base two unless specified.  $i(x)$  of a certain realization, so a certain outcome of a random experiment, encodes the amount of *surprise* that one has upon the gaining said information. The less probable a certain outcome/realization is, the greater information content  $i(x)$  holds. The logarithm is a good choice to model this surprise because it is additive. That is, given two independent random experiments involving random variable  $X$ ,

$$i(x_1, x_2) = -\log(p_{X,X}(x_1, x_2)) = -\log(p_X(x_1)p_X(x_2)) = i(x_1) + i(x_2) \quad (2)$$

$i(x)$ , however, is dependent on certain realization  $x$  of a random variable  $X$ ; it is inept at representing the expected level of surprise of  $X$ . Hence, we define the entropy  $H(X)$  to represent the level of surprise of a random variable  $X$  – as Wilde puts it, *the entropy is the expected information content of a random variable*. We define  $H(X)$  as follows:

$$H(X) \equiv \mathbb{E}_X\{i(X)\} \quad (3)$$

**Definition 2.1. (Entropy)** The entropy of a discrete random variable  $X$  with probability distribution  $p_X(x)$  is

$$H(X) \equiv -\sum_x p_X(x) \log(p_X(x)) \quad (4)$$

$0 \cdot \log(0) = 0$  for conventional reasons. This definition is fitting because if a random variable  $X$  had a degenerate probability distribution (all realizations  $x$  of  $X$  have probability zero besides a singular realization), the amount of information one would learn from the outcome of an experiment modeled by  $X$  would be 0. We gain no meaningful information (level of surprise, so to speak) if we know something is guaranteed to happen. Another interesting thing to consider preliminarily is if  $p_X(x)$  is uniform. What does the entropy of a random variable  $X$  with uniform probability distribution tell us then?

## 2.1 Binary Entropy Function

When the random variable we are dealing with is Bernoulli (analogous to a coin flip) with probability density  $p_X(0) = p$ ,  $p_X(1) = 1 - p$ , the entropy function is as follows:

**Definition 2.2. (Binary Entropy)** The binary entropy of  $p \in [0, 1]$  is

$$h_2(p) \equiv -p \log(p) - (1 - p) \log(1 - p) \quad (5)$$

## 2.2 Properties of Entropy

**Proposition 2.3. (Non-Negativity)** The entropy  $H(X)$  is non-negative for any discrete random variable  $X$  with probability density  $p_X(x)$ :

$$H(X) \geq 0 \quad (6)$$

*informal proof.* This is rather intuitive; we always learn something when provided with a certain realization  $x$  of  $X$ . An outcome of an experiment cannot give us a negative amount of information.

**Proposition 2.4. (Concavity)** The entropy  $H(X)$  is concave in the probability density  $p_X(x)$ .

**Proposition 2.5. (Permutation Invariance)** The entropy is invariant with respect to permutations of the realizations of the random variable  $X$

**Proposition 2.6. (Minimum Value)** The entropy  $H(X)$  vanishes if and only if  $X$  is a deterministic variable.

Sometimes we are completely blind; we have no information about even the possible values of a system, but we need a probability distribution to describe it. The most reasonable thing to do is to come up with some way of making a *conservative* estimate or guess at the actual form of the distribution. To elaborate, we want to be able to gain insight into the system while minimizing the odds of us making overly favorable assumptions. That is, our assumption would be most valid by shouldering, carrying with us, so to speak, a maximal amount of uncertainty.

**Proposition 2.7. (Maximum Value)** The maximum value of the entropy  $H(X)$  for a random variable  $X$  taking values in an alphabet  $\chi$  is  $\log(|\chi|)$ :

$$H(X) \leq \log(|\chi|) \quad (7)$$

The inequality is saturated if and only if  $X$  is a uniform random variable on  $\chi$ .

### 3 Conditional Entropy

Let's say Alice has access to a random variable  $X$ , while Bob has access to some other random variable, say  $Y$ . It is not unreasonable to claim that unless  $X$  and  $Y$  are independent, Alice possess some information about  $Y$  in the form of  $X$ ; in the same vein Bob would possess some information about Alice's  $X$  in the form of  $Y$ . We now build conditional entropy from the ground up using the notion of information content.

$$i(x|y) = -\log(p_{X|Y}(x|y)) \quad (8)$$

is the information content of a realization  $x$  conditioned on a realization  $y$ . Naturally,  $H(X|Y = y)$  of a random variable  $X$  conditioned on a particular realization of  $y$  of a random variable  $Y$  is the expected conditional information content with respect to  $X|Y = y$  – how much information are we expected to earn (in bits) about an experiment modeled by  $X$  given a certain outcome  $y$ ?<sup>1</sup>

$$\begin{aligned} H(X|Y = y) &\equiv \mathbb{E}_{X|Y=y} \{i(X|y)\} \\ &= - \sum_x p_{X|Y}(x|y) \log(p_{X|Y}(x|y)) \end{aligned} \quad (9)$$

**Definition 3.1. (Conditional Entropy)** Let  $X$  and  $Y$  be discrete random variables with joint probability distribution  $p_{X,Y}(x, y)$ . The conditional entropy  $H(X|Y)$  is the expected conditional information content, where the expectation is with respect to both  $X$  and  $Y$ :

$$\begin{aligned} H(X|Y) &\equiv \mathbb{E}_{X,Y} \{i(X|Y)\} \\ &= \sum_y p_Y(y) H(X|Y = y) \\ &= - \sum_y p_Y(y) \sum_x p_{X|Y}(x|y) \log(p_{X|Y}(x|y)) \\ &= - \sum_{x,y} p_{X,Y}(x, y) \log(p_{X|Y}(x|y)) \end{aligned} \quad (10)$$

Wilde gives the following interpretation: *Suppose that Alice possesses random variable  $X$  and Bob possesses random variable  $Y$ . The conditional entropy  $H(X|Y)$  is the amount of uncertainty that Bob has about  $X$  given that he already possesses  $Y$ .* A logical continuation of this observation is that having access to another random variable should never increase our uncertainty regarding another different random variable.

<sup>1</sup>don't know why an arrow is appearing instead of  $y$ . Will troubleshoot later!

**Theorem 3.2. (Conditioning Does Not Increase Entropy)** The entropy  $H(X)$  is greater than or equal to the conditional entropy  $H(X|Y)$ :

$$H(X) \geq H(X|Y) \quad (11)$$

and equality occurs if and only if  $X$  and  $Y$  are independent random variables.

The last sentence makes intuitive sense – if we have access to ancilla system that is completely independent (shares no mutual information with) of a probe system, information about the additional ancilla system will never shed light on our probe system.

## 4 Joint Entropy

**Definition 4.1. (Joint Entropy)** Let  $X$  and  $Y$  be discrete random variables with joint probability distribution  $p_{X,Y}(x, y)$ . The joint entropy  $H(X, Y)$  is defined as

$$H(X, Y) \equiv \mathbb{E}_{X,Y}\{i(X, Y)\} = - \sum_{x,y} p_{X,Y}(x, y) \log(p_{X,Y}(x, y)) \quad (12)$$

### 4.1 Joint Entropy Exercises

#### 4.1.1 Exercise 10.3.1

Verify that  $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$

*proof.* We will show that  $H(X, Y) = H(X) + H(Y|X)$  as follows:

$$H(X) + H(Y|X) = - \sum_x p_X(x) \log(p_X(x)) - \sum_{x,y} p_{X,Y}(x, y) \log(p_{Y|X}(y|x)) \quad (13)$$

Recall that  $p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$ . Substituting and making use of the properties of the logarithm, the right-hand summation of (1) is

$$- \sum_{x,y} p_{X,Y}(x, y) \log(p_{X,Y}(x, y)) - p_{X,Y}(x, y) \log(p_X(x)) \quad (14)$$

distributing the double sum inwards and marginalizing gives us

$$\sum_{x,y} p_{X,Y}(x, y) \log(p_X(x)) = \sum_x p_X(x) \log(p_X(x)) \quad (15)$$

substituting and cleaning up terms gives us that

$$H(X) + H(Y|X) = H(X) + (H(X, Y) - H(X)) = H(X, Y), \quad (16)$$

as desired.  $H(X, Y) = H(Y) + H(X|Y)$  follows via the same exact steps.

#### 4.1.2 Exercise 10.3.2

Extend the result of the previous exercise and to prove the following chain rule for entropy:

$$H(X_1, \dots, X_n) = H(X_1) + H(X_2|X_1) + \dots + H(X_n|X_{n-1}, \dots, X_1) \quad (17)$$

*proof.* We will proceed via induction on  $n$ . Ignoring the trivial case where  $n = 1$ , we omit a proof of the base case when  $n = 2$  as we can refer directly to the proof given above. Let us suppose that the following equality holds for all  $k \leq n - 1$ :  $H(X_1, \dots, X_{n-1}) = H(X_1) + H(X_2|X_1) + \dots + H(X_{n-1}|X_{n-2}, \dots, X_1)$ .<sup>a</sup> Now, notice that

$$H(X_1, \dots, X_n) = - \sum_{x_1, \dots, x_n} p_{X_1, \dots, X_n}(x_1, \dots, x_n) \log(p_{X_1, \dots, X_n}(x_1, \dots, x_n)) \quad (18)$$

Now we refer to the<sup>b</sup> following:

$$p_{X_n|X_{n-1}, \dots, X_1}(x_n|x_{n-1}, \dots, x_1) \cdot p_{X_1, \dots, X_{n-1}}(x_1, \dots, x_{n-1}) = p_{X_1, \dots, X_n}(x_1, \dots, x_n) \quad (19)$$

substituting into the logarithm in  $H(X_1, \dots, X_n)$ , expanding, and then separating into two terms we get the following two equalities:

$$- \sum_{x_1, \dots, x_n} p_{X_1, \dots, X_n}(x_1, \dots, x_n) \log(p_{X_n|X_{n-1}, \dots, X_1}(x_n|x_{n-1}, \dots, x_1)) = H(X_n|X_{n-1}, \dots, X_1) \quad (20)$$

and

$$- \sum_{x_1, \dots, x_{n-1}} \log(p_{X_1, \dots, X_{n-1}}(x_1, \dots, x_{n-1})) \sum_{x_n} p_{X_1, \dots, X_n}(x_1, \dots, x_n) = H(X_{n-1}) \quad (21)$$

Therefore, we have split

$$H(X_1, \dots, X_n) = H(X_n|X_{n-1}, \dots, X_1) + H(X_{n-1}) \quad (22)$$

applying the inductive hypothesis on  $H(X_{n-1})$  gives us that

$$H(X_1, \dots, X_n) = H(X_1) + H(X_2|X_1) + \dots + H(X_n|X_{n-1}, \dots, X_1), \quad (23)$$

as desired.

<sup>a</sup>The burden of proof is to now show  $H(X_1, \dots, X_n) = H(X_1) + H(X_2|X_1) + \dots + H(X_n|X_{n-1}, \dots, X_1)$ .

<sup>b</sup><https://stats.stackexchange.com/questions/258379/why-is-pa-bc-pbc-pab-c>

#### 4.1.3 Exercise 10.3.3

Prove that entropy is subadditive:

$$H(X_1, \dots, X_n) \leq \sum_{i=1}^n H(X_i) \quad (24)$$

*proof.* We make use of the previous exercise and the fact that *conditioning does not increase entropy*.

$$H(X_1, \dots, X_n) = H(X_1) + H(X_2|X_1) + \dots + H(X_n|X_{n-1}, \dots, X_1) \quad (25)$$

Refer now to the following theorem:

**Theorem 4.2.** The entropy  $H(X)$  is greater than or equal to the conditional entropy  $H(X|Y)$

$$H(X) \geq H(X|Y) \quad (26)$$

Repeated application of the above theorem gives us  $H(X_1, \dots, X_n) \leq \sum_{i=1}^n H(X_i)$ , as desired.

#### 4.1.4 Exercise 10.3.4

Prove that entropy is additive when the random variables  $X_1, X_2, \dots, X_n$  are independent:

$$H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i) \quad (27)$$

*proof.*

$$H(X_1, \dots, X_n) = - \sum_{x_1, \dots, x_n} p_{X_1, \dots, X_n}(x_1, \dots, x_n) \log(p_{X_1, \dots, X_n}(x_1, \dots, x_n)) \quad (28)$$

Since the random variables are independent,

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = p_{X_1}(x_1) \cdot p_{X_2}(x_2) \cdot \dots \cdot p_{X_n}(x_n) \quad (29)$$

Armed with this fact, we can substitute this expression into the logarithm, use the additive property of the logarithm to expand the summation, and finally, distribute the summations to marginalize the probabilities, giving us the equality  $H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i)$ , as desired.

*we will omit from explicitly showing the computations as they are essentially repeated applications of the techniques employed first two problems of this solution sheet.*

## 5 Mutual Information

Let us now return our attention back to Alice and Bob. Say Alice has random variable  $X$  and Bob has random variable  $Y$ . We now present an entropic measure of the mutual information that both Alice and Bob possess.



**Definition 5.1. (Mutual Information)** Let  $X$  and  $Y$  be discrete random variables with joint probability distribution  $p_{X,Y}(x, y)$ . The mutual information  $I(X; Y)$  is the marginal entropy  $H(X)$  less the conditional entropy  $H(X|Y)$

$$I(X; Y) \equiv H(X) - H(X|Y) \quad (30)$$

Mutual information quantifies how much having knowledge about one random variable (or system) reduces the uncertainty about the other random variable (system). Since Bob has access to  $Y$ , Bob has an uncertainty that reduced by  $H(X|Y)$ , since the aforementioned conditional entropy quantifies the amount of uncertainty Bob has about Alice's  $X$  given his access to  $Y$ .

## 5.1 Mutual Information Exercises

### 5.1.1 Exercise 10.4.1 (problem statement omitted due to spacing reasons)

*proof.* Via the definition of mutual information we get that

$$I(X; Y) = H(X) - H(X|Y) \quad (31)$$

since we need to figure out a way to get rid of  $H(X)$  in the expression above, let us leave it there for now and get to manipulating the scarier looking  $H(X|Y)$ .

$$H(X|Y) = - \sum_{x,y} p_{X,Y}(x, y) \log(p_{X|Y}(x|y)) \quad (32)$$

Recall that  $p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_X(y)}$ . Substituting <sup>a</sup>,

$$H(X|Y) = - \sum_{x,y} p_{X,Y}(x, y) \log\left(\frac{p_{X,Y}(x, y)}{p_X(y)}\right) \quad (33)$$

Using the subtractive properties of the logarithm and plugging and chugging,

$$= - \sum_{x,y} p_{X,Y}(x, y) \log(p_{X,Y}(x, y)) - \sum_y \log(p_X(y)) \sum_x p_{X,Y}(x, y) \quad (34)$$

marginalizing gives us that

$$H(X|Y) = H(X, Y) - H(Y) \quad (35)$$

Then, substituting this relation into the very first equation of the proof,

$$I(X; Y) = H(Y) - H(Y|X) = I(Y; X) \quad (36)$$

as desired.

---

<sup>a</sup>this technique where we substitute only in the logarithm is essential because it allows us to marginalize the joint probability later. We also used this technique in some of the solutions above

The symmetry relation proven above lets us know that given Alice's possession of  $X$ , her uncertainty about  $H(Y)$  is reduced by  $H(Y|X)$ , the amount of uncertainty Alice has about Bob's  $Y$  given her access to  $X$ . We can also express  $I(X;Y)$  in terms of the respective joint and marginal probability density functions  $p_{X,Y}(x,y)$  and  $p_X(x)$  and  $p_Y(y)$ :

$$I(X;Y) = \sum_{X,Y} p_{X,Y}(x,y) \log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right) \quad (37)$$

The expression above tells us immediately that the mutual information that Alice and Bob share if  $X$  and  $Y$  are independent is zero. We finish with an intuitive theorem that bounds mutual information above zero, which honestly speaking, follows directly from the fact that entropy is non-negative and that conditional entropy is bounded below entropy.

**Theorem 5.2.** The mutual information  $I(X;Y)$  is non-negative for any random variables  $X$  and  $Y$ :

$$I(X;Y) \geq 0, \quad (38)$$

and  $I(X;Y)$  is zero if and only if  $X$  and  $Y$  are statistically independent random variables.

## 6 Relative Entropy

The relative entropy is best seen as a *pseudo-metric* or *pseudo-distance* that can be used to quantify how far apart a probability density function  $p(x)$  is from another probability density function  $q(x)$ . We will see why we attach *pseudo* shortly.

**Definition 6.1. (support)** Let  $\chi$  denote a finite set. The support of a function  $f : \chi \rightarrow \mathbb{R}$  is equal to the subset of  $\chi$  that takes non-zero values under  $f$ :

$$\text{supp}(f) \equiv \{x : f(x) \neq 0\}. \quad (39)$$

**Definition 6.2. (Relative Entropy)** let  $p$  be a probability distribution defined on the alphabet  $\chi$ , and let  $q : \chi \rightarrow [0, \infty)$ . The relative entropy  $D(p||q)$  is defined as follows:

$$D(p||q) \equiv \begin{cases} \sum_x p(x) \log(p(x)/q(x)) & \text{if } \text{supp}(p) \subseteq \text{supp}(q) \\ \infty & \text{else} \end{cases} \quad (40)$$

The properties of the logarithm are symmetry breaking; we cannot interchange  $p(x)$  and  $q(x)$  without modifying the value of the relative entropy. Hence, we cannot consider it a distance, so we'll use *pseudo* since it still does capture, perhaps a little vaguely, the notion of distances between probability distributions. A more mathematical treatment of the relative entropy can be looked into by the reader by looking into the *Kullback Leibler Divergence*, which I believe is used interchangeably with the term relative entropy (I may be mistaken). Wilde provides a great explanation/interpretation of relative entropy as follows: *Suppose that an information source generates a random variable  $X$  according to the density*

$p(x)$ . Suppose further that Alice mistakenly assumes that the probability density function is  $q(x)$  and codes according to this density. Then the relative entropy quantifies the inefficiency that Alice incurs when she codes according to the mistaken probability density.

## 7 Conditional Mutual Information

**Definition 7.1. (Conditional Mutual Information)** Let  $X, Y, Z$  be discrete random variables. The conditional mutual information is defined as follows:

$$I(X; Y|Z) \equiv H(Y|Z) - H(Y|X, Z) \quad (41)$$

$$I(X; Y|Z) = H(X|Z) - H(X|Y, Z) \quad (42)$$

$$I(X; Y|Z) = H(X|Z) + H(Y|Z) - H(X, Y|Z) \quad (43)$$

**Theorem 7.2. (Strong Subadditivity)** The conditional mutual information  $I(X; Y|Z)$  is non-negative:

$$I(X; Y|Z) \geq 0, \quad (44)$$

and the inequality is saturated if and only if  $X - Y - Z$  is a Markov chain (i.e., if  $p_{X,Y|Z}(x, y|z) = p_{X|Z}(x|z)p_{Y|Z}(y|z)$ )

## 8 Entropy Inequalities

### 8.1 Non-Negativity of Relative Entropy

**Theorem 8.1. (Non-Negativity of Relative Entropy)** Let  $p(x)$  be a probability distribution over the alphabet  $\chi$  and let  $q : \chi \rightarrow [0, 1]$  be a function such that  $\sum_x q(x) \leq 1$ . Then the relative entropy  $D(p||q)$  is non-negative:

$$D(p||q) \geq 0, \quad (45)$$

and  $D(p||q) = 0$  if and only if  $p = q$ .

**proof.** First, let us posit that  $\text{supp}(p) \not\subseteq \text{supp}(q)$ . Then, the relative entropy jumps to infinity; it is trivially bounded above zero. Now, let us suppose that  $\text{supp}(p) \subseteq \text{supp}(q)$ . Note that  $\ln(x) \leq 1 - x$  for all  $x \geq 0$  and the inequality saturates only when  $x = 1$ . We will omit a rigorous proof of this in this text as this is isn't an analysis class <sup>2</sup>. We mention this now because we will make good use of it in the chain of equalities to come.

$$D(p||q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right) \quad (46)$$

<sup>2</sup>Maybe this is heinous for a math major to say but even a graphical argument would suffice at least for the purpose of this document

Recall that  $\log(\frac{p(x)}{q(x)}) = \frac{\ln(p(x)/q(x))}{\ln(2)}$  and rearranging the terms a bit gives us that  $\ln(q(x)/p(x)) = -\frac{1}{\ln(2)} \log(\frac{p(x)}{q(x)})$ . Then,

$$D(p||q) = -\frac{1}{\ln(2)} \sum_x p(x) \ln\left(\frac{q(x)}{p(x)}\right) \quad (47)$$

Recall now that  $\ln(x) \leq x - 1$  implies that  $-\ln(x) \geq 1 - x$ . Then,

$$D(p||q) \geq \frac{1}{\ln(2)} \sum_x p(x) \left(1 - \frac{q(x)}{p(x)}\right) \quad (48)$$

Distributing and using the fact that  $\sum_x p(x)$  and  $\sum_x q(x)$  sum to one, we get that

$$D(p||q) \geq \frac{1}{\ln(2)} \sum_x p(x) \left(1 - \frac{q(x)}{p(x)}\right) = 0. \quad (49)$$

Now, if  $p = q$ ,  $D(p||q)$  follows trivially. Suppose now then that  $D(p||q) = 0$ . This means that  $\text{supp}(p) \subseteq \text{supp}(q)$  by construction. Recall once more that  $\ln(x) \leq x - 1$  and that this bound is only saturated if  $x = 1$ . This, paired with the fact that  $\ln(\frac{q(x)}{p(x)}) = 0$  implies that  $p = q$ , as desired. We now give a proof of 2.7, which made the claim that  $H(X) \leq \log(|\chi|)$ .

**proof.** Using the *non-negativity* of the relative entropy, we proceed as follows:

$$0 \leq D(p_X(x)||\{\frac{1}{|\chi|}\}) \quad (50)$$

$$D(p_X(x)||\{\frac{1}{|\chi|}\}) = \sum_x p_X(x) \log\left(\frac{p(x)}{\frac{1}{|\chi|}}\right) \quad (51)$$

Expanding using the properties of the logarithm we get that

$$0 \leq \sum_x p_X(x) \log(p_X(x)) + \sum_x p_X(x) \log |\chi| \quad (52)$$

Via the definition of entropy we get that,

$$0 \leq -H(X) + \sum_x p_X(x) \log |\chi| \quad (53)$$

which implies that (moving terms around and using the fact that the probabilities sum to one)

$$H(X) \leq \log |\chi|, \quad (54)$$

as desired.

## 8.2 Data-Processing Inequality

## 9 Near Saturation of Entropy Inequalities

## References